R E P O R T R E S U M E S

ED 017 006 24 CG 001 556

EFFECT OF ERROR OF MEASUREMENT ON THE POWER OF STATISTICAL TESTS. FINAL REPORT.

BY- CLEARY, T. A. LINN, ROBERT L.
EDUCATIONAL TESTING SERVICE, PRINCETON, N.J.

REPORT NUMBER BR-6-8574 PUB DATE SEP 67
GRANT OEG-1-7-06874-2632
EDRS PRICE MF-$0.25 HC-$2.08 50P.

DESCRIPTORS- *STATISTICAL ANALYSIS, MENTAL TESTS, *RELIABILITY, TEST CONSTRUCTION, *TESTS OF SIGNIFICANCE, ANALYSIS OF VARIANCE, *MEASUREMENT TECHNIQUES,

THE PURPOSE OF THIS RESEARCH WAS TO STUDY THE EFFECT OF ERROR OF MEASUREMENT UPON THE POWER OF STATISTICAL TESTS. ATTENTION WAS FOCUSED ON THE F-TEST OF THE SINGLE FACTOR ANALYSIS OF VARIANCE. FORMULAS WERE DERIVED TO SHOW THE RELATIONSHIP BETWEEN THE NONCENTRALITY PARAMETERS FOR ANALYSES USING TRUE SCORES AND THOSE USING OBSERVED SCORES. THE EFFECT OF THE RELIABILITY OF THE MEASUREMENT AND THE SAMPLE SIZE WERE THUS DEMONSTRATED. THE ASSUMPTIONS OF CLASSICAL TEST THEORY WERE USED TO DEVELOP FORMULAS RELATING TEST LENGTH TO THE NONCENTRALITY PARAMETERS. THREE METHODS OF ESTIMATING POWER FOR DIFFERENT CONDITIONS OF SAMPLE SIZE AND TEST LENGTH WERE STUDIED. THE COST OF AN EXPERIMENT WAS ANALYZED IN TERMS OF A FIXED COST PER SUBJECT AND A VARIABLE COST DEPENDENT UPON TEST LENGTH. COMPUTER PROGRAMS WERE WRITTEN TO USE THE LEAST SQUARES APPROXIMATION AND THE APPROXIMATION BASED ON PATNAIK TO ESTIMATE THE POWER UNDER ALL PERMISSIBLE ALLOCATIONS OF RESOURCES TO SAMPLE SIZE AND TEST LENGTH. THE PROGRAM RESULTS INDICATE WHICH OF THE PERMISSIBLE ALLOCATIONS WILL RESULT IN MAXIMUM POWER. TO DEMONSTRATE EMPIRICALLY THE EFFECT OF ERROR OF MEASUREMENT ON THE POWER OF STATISTICAL TESTS, SAMPLES OF PERSONS AND ITEMS WERE RANDOMLY DRAWN FROM A LARGE POOL OF DATA. TESTS OF 10, 20, AND 40 RANDOMLY DRAWN ITEMS WERE SCORED FOR SAMPLES WITH FOUR AND EIGHT PERSONS PER GROUP. THE EXPECTED TRENDS WERE PRESENT BUT NOT DEFINITIVE. (AUTHOR)

ED017006

FINAL REPORT
Project No. 6-8574-24
Contract No. OEG-1-7-068574-2632

EFFECT OF ERROR OF MEASUREMENT ON THE POWER OF STATISTICAL TESTS

September 1967

U.S. DEPARTMENT OF HEALTH,
EDUCATION, AND WELFARE

Office of Education
Bureau of Research

CG 001 556

Effect of Error of Measurement on the Power of Statistical Tests

Project No. 6-8574
Contract No. OEG-1-7-068574-2632

T. Anne Cleary* and Robert L. Linn

September 1967

The research reported herein was performed pursuant to a grant with the Office of Education, U.S. Department of Health, Education, and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education position or policy.

Educational Testing Service

Princeton, New Jersey

*The services of Dr. Cleary were subcontracted with the University of Wisconsin.

Contents

	Page
Introduction	1
Problem	1
Purpose	2
Part I: Theoretical Development	2
Test Theory	2
Statistical Tests	4
The Power Function	8
Cost of an Experiment	15
Allocation of Resources	16
Conclusions	18
Part II: Empirical Demonstration	21
Purpose	21
Method	21
Results	22
Discussion	27
Conclusions	28
Summary	28
References	30
Appendix A	A-1
Appendix B	B-1

Introduction

Problem

Discussions of the power of statistical tests can be found in almost all basic statistics books. The power function, which gives the probability of rejecting a hypothesis, depends upon the differences expected in random samples from the same population, that is, upon sampling error. Implicit in the usual discussion of power is the assumption that the observations are errorless or "true" measurements. Sampling error rather than measurement error is considered.

The test theory literature, on the other hand, is concerned primarily with the error of measurement (4). Observations are considered fallible and repeated measures of the same object are expected to vary about the "true" measurement, the expected value of the repeated measures.

Sutcliffe (10) has attempted to consider the two types of error simultaneously. Sutcliffe elaborated the implications of measurement error for the F test of differences between means and demonstrates how measurement error decreases the sensitivity of a test of significance. More specifically, Sutcliffe compared the ratios of the expected mean square between groups to the expected mean square within groups for a single factor analysis of variance in two cases: the case of no measurement error and the case where observed scores were assumed to include measurement error as defined in classical test theory. Sutcliffe showed that the power of the test is always greater for the error-free case.

Lord (6) has given extensive consideration to the implications of an item sampling model for mental test theory. Lord has shown that item sampling methods can improve the efficiency of the experimental design of a study particularly one concerned with group means.

> (i) If only a limited amount of time can be demanded of each research subject, the total amount of information obtained from a given number of subjects may be greatly increased by item sampling. (ii) If a test can be administered to only one examinee at a time, the examiner's time may be the limiting factor; more information about a group of examinees may be obtained by giving a few items to each examinee instead of giving the entire test to just a few examinees. (iii) With certain tests, scoring costs may be the limiting factor; in this case, it would be better to score a few items from the answer sheet of each examinee than to score all items on the answer sheets of a few examinees. (6, p. 23)

The item-sampling model has strong advantages in many group-comparison situations such as frequently occur in the evaluations of educational programs. However, practical administrative considerations such as the need for common instructions and testing time, the economy of being able to use a single scoring key, and the fact that test data must frequently serve several purposes, often make it desirable to administer the same test to all examinees. In such situations, one is faced with the problem of deciding whether it is more efficient to improve the sensitivity of a planned statistical test by increasing the number of examinees or by increasing the test length as a means of reducing the error of measurement.

Overall and Dalal (7) discussed the problem of choosing a research design which maximizes power relative to cost. They concluded that no matter how unreliable the measurement, it is better to use more subjects and obtain a single measurement per subject than to obtain several measures on each of fewer subjects. As Overall and Dalal pointed out, the above conclusion is based on the assumption that there is a fixed cost per measurement unit, and this cost is the same whether the units are obtained for the same subject or different subjects.

Purpose

The purpose of this research was to develop, from the assumptions of classical test theory, formulas demonstrating the effect of error of measurement on the power of some commonly used statistical tests. An important aspect of the research was the development of a procedure that would enable the educational researcher to estimate whether an attempt to reduce measurement error by increasing accuracy of observations or to reduce sampling error by increasing the number of observations would be the more effective strategy. The implications that various assumptions concerning the fixed and variable costs of testing have for the choice of a strategy were investigated also. Since the assumptions of classical theory cannot be expected to hold exactly in real data, the effects on statistical tests of increasing reliability and the number of observations were demonstrated empirically.

Part 1: Theoretical Development

Test Theory

In classical test theory, it is assumed that an observation, X_i, for individual i is equal to his true score, T_i, plus an error score, E_i:

(1) $$X_i = T_i + E_i \ ,$$

where the expected value of E equals zero $(\mathcal{E}(E) = 0)$, the variance of E equals σ_E^2, and the covariance of T with E, σ_{TE}, equals zero (4).

Given these assumptions, it can be shown that:

(2) $$\mathcal{E}(X) = \mathcal{E}(T)$$

and

(3) $$\sigma_X^2 = \sigma_T^2 + \sigma_E^2 \ ,$$

where σ_X^2 is the variance of X and σ_T^2 is the variance of T.

If ρ is the reliability of measurement X, then the variance of the error can be written:

(4) $$\sigma_E^2 = \sigma_X^2 (1 - \rho) = \sigma_T^2 \frac{(1-\rho)}{\rho} \ .$$

If a test is lengthened by combining K unit-length parallel tests, the relationships between the parameters of the unit length test and those of the lengthened test are well known:

(5) $$\sigma_{T_K}^2 = K^2 \sigma_{T_1}^2$$

(6) $$\sigma_{X_K}^2 = [K + K(K-1) \rho_1] \sigma_{X_1}^2$$

(7) $$\sigma_{E_K}^2 = K\sigma_{E_1}^2$$

and

(8) $$\rho_K = \frac{K\rho_1}{1 + (K-1) \rho_1} \ ,$$

where the subscript K denotes the lengthened test and the subscript 1 denotes the unit length test. From the above formulas, it is apparent that, if K is larger than one, the three variances increase with K : the increase is greatest for the variance of the true scores, least for the variance of the error. The change in the relative sizes of the variances is reflected in the change in the reliability: as K increases, the reliability increases.

Statistical Tests

In the derivation and interpretation of statistical tests, the observations are generally considered to be free of error of measurement, that is, in the language of test theory, the observations are true scores. The application of statistical tests to observed scores subject to error of measurement is in no sense incorrect or even necessarily inappropriate: the assumptions of the statistical tests may be satisfied by the observed scores. However, if the hypotheses are formulated in terms of true scores and tested with observed scores, the noncentrality parameter and therefore the power can be quite different from what would be expected with true scores. Failure to reject the null hypothesis with observed scores is not equivalent to a failure to reject the null hypothesis with true scores.

Perhaps, one of the most commonly used statistical tests in educational research is the F test of the analysis of variance. In addition to being commonly used, it is well known that the F test with one and ν_2 degrees of freedom is equivalent to the two-tailed t test. If ν_2 approaches infinity, the F distribution approaches a chi-square distribution.

Consider a single-factor analysis of variance. The model for this analysis is

$$T_{ig} = M + A_g + B_{ig} \tag{9}$$

$$g = 1, \ldots, G$$

$$i = 1, \ldots, n$$

where

T_{ig} is the true score for individual i in group g;
M is the population true-score mean,
A_g is the component of the true score which is due to the effect of treatment g, and
B_{ig} is the deviation of an individual's score from the group mean, the error of analysis-of-variance model.

The B_{ig} are assumed to be independently and normally distributed with expected value of zero and common variance σ_B^2. Over all possible treatments, g, the sum of the A_g is zero and the variance is σ_A^2. Table 1 presents the expected mean squares for this model.

TABLE 1

Expected Mean Squares for a Single-Factor Analysis of Variance of True Scores

Source	Degrees of Freedom	$\mathcal{E}$(MS)
Between	G-1	$n\sigma_A^2 + \sigma_B^2$
Within	G(n-1)	σ_B^2
Total	Gn-1	

If the null hypothesis of no difference between treatments ($\sigma_A^2 = 0$) is true, the test statistic (the ratio of the mean square between groups to the mean square within groups) is distributed as F with (G-1) and G(n-1) degrees of freedom. If the null hypothesis is not true, the test statistic is distributed as a noncentral F with the same degrees of freedom and noncentrality parameter:

(10) $$\lambda_T = \frac{n\sigma_A^2}{\sigma_B^2} \quad .$$

If observed scores rather than true scores are used in the analysis, the model is

(11) $$X_{ig} = M + A_g + B_{ig} + E_{ig}$$

where

X_{ig} is the observed score for individual i in group g,
E_{ig} is the measurement error for individual i in group g, and
M, A_g, and B_{ig} are the same as in the true score model.

Within each group g, the measurement error, E_{ig}, is assumed to have a normal distribution with expected value of zero and variance, σ_E^2. The expected mean squares for this analysis are shown in Table 2.

TABLE 2

Expected Mean Squares for a Single-Factor Analysis of Variance of Observed Scores

Source	Degrees of Freedom	$\mathcal{E}$(MS)
Between	G-1	$n\sigma_A^2 + \sigma_B^2 + \sigma_E^2$
Within	G(n-1)	$\sigma_B^2 + \sigma_E^2$
Total	Gn-1	

If the null hypothesis $(\sigma_A^2 = 0)$ is true, the test statistic has the same distribution as in the error-free case. However, if the null hypothesis is false, the test statistic is distributed as noncentral F with the same degrees of freedom but with noncentrality parameter,

$$\lambda_X = \frac{n\sigma_A^2}{\sigma_B^2 + \sigma_E^2} \quad . \tag{12}$$

For σ_E^2 greater than zero, the noncentrality parameter for the observed score analysis, λ_X, will be smaller than the noncentrality parameter for the true score analysis, λ_T. Since power for the test with given degrees of freedom is a nondecreasing function of the noncentrality parameter, the power for the true-score analysis is always greater than the power for the observed score analysis.

For fixed n, the relationship between power and error of measurement can be seen by noting that the ratio of the mean squares divided by their expected values,

$$\frac{MS_{Between}}{MS_{Within}} \cdot \frac{\sigma_B^2 + \sigma_E^2}{n\sigma_A^2 + \sigma_B^2 + \sigma_E^2},$$

is distributed as Central F. Power can then be expressed as

$$\Pr \left\{ \frac{MS_{Between}}{MS_{Within}} \quad F \cdot \frac{1}{1+\lambda} \right\} .$$

Clearly, the larger λ, the smaller the term to the right of the inequality sign and the greater the power. As mentioned earlier this result was obtained previously by Sutcliffe (10).

The relationship between power and λ is, of course, dependent upon the degrees of freedom. For fixed degrees of freedom the power is a negatively accelerated function of λ. As the degrees of freedom in the denominator (or number of persons per cell) increases, the initial slope increases and also the rate of negative acceleration.

The noncentrality parameter can be usefully expressed in terms of ρ, the reliability of the measure and the variances of the true score components. From formulas 4 and 12

(13) $$\lambda_X = \frac{\rho n \sigma_A^2}{\sigma_B^2 + (1-\rho)\sigma_A^2} ,$$

since

(14) $$\sigma_T^2 = \sigma_A^2 + \sigma_B^2 .$$

The relationship between the noncentrality parameters for the observed score and true-score analyses can be seen by substituting formula 10 into 13:

(15) $$\lambda_X = \frac{n\rho \lambda_T}{n + (1-\rho)\lambda_T} .$$

For fixed λ_T, and n, λ_X is a positively accelerated function of ρ, the reliability of the scores: as ρ increases by equal units from zero, the increase in λ_X is at first quite small but each successive increase in ρ results in a slightly larger increase in λ_X. The rate of positive acceleration increases to λ_T.

If the test length is increased by a factor of K, the reliability and therefore the noncentrality parameter will increase. If ρ_1 denotes the reliability of the unit length test, the observed scores noncentrality parameter can be expressed:

(16) $$\lambda_X = \frac{nK\rho_1\lambda_T}{nK\rho_1 + (1-\rho_1)(\lambda_T + n)} .$$

The effect of n and K on the noncentrality parameter, λ_X, can be seen more clearly if equation 16 is expressed in terms of ϕ_T^2 where

(17) $$\phi_T^2 = \frac{\sigma_A^2}{\sigma_B^2} .$$

Thus, $\lambda_T = n\phi_T^2$ and:

(18) $$\lambda_X = \frac{nK\rho_1\phi_T^2}{K\rho_1 + (1-\rho)(\phi_T^2 + 1)} .$$

The noncentrality parameter, λ_X, is a strictly increasing function of both K and n. However, the effect of increasing n is relatively greater than the effect of increasing K since K influences both the numerator, and the denominator whereas n influences only the numerator. In addition, the effect of n upon power is increased by the change in degrees of freedom.

The Power Function

The power function for a statistical test gives the probability that the null hypothesis will be rejected given different alternative values of the parameter. To determine the power of the F test of the analysis of variance, one needs to determine the proportion of the area of the noncentral F distribution that falls in the critical region. In the single factor analysis of variance, the test statistic, F_o, is:

(19) $$F_o = \frac{MS_{Between}}{MS_{Within}}$$

and the critical region is defined by

$$F_o > F_\alpha$$

where α is the significance level of the test. The power function for a given λ is then given by

(20) $$\text{Power} = \int_{F_\alpha}^{\infty} p_{\nu_1\nu_2}(F'|\lambda)\, dF'$$

where F_α is the percentage point of the F distribution with degrees of freedom ν_1, ν_2 and the integration is over the density function of the noncentral F distribution, F', with ν_1 and ν_2 degrees of freedom and noncentrality parameter λ.

The evaluation of the power function is not simple. Methods of evaluating the probability integral have been worked out by Wishart (12) and Tang (11), but the amount of labor involved generally limits consideration to a few alternative hypotheses. Several authors have presented power function curves (2, 3, 5, 8, 9). These curves enable one to determine quickly, if approximately, the power for a limited number of sets of degrees of freedom and noncentrality parameters. The most relevant of these charts for the design of experiments are those of Feldt and Mahmoud (2) which present curves of constant power, for power equal to .5, .7, .9, as a function of n, the number of persons per cell, and ϕ, the noncentrality parameter. The charts are designed to permit the specification of sample size for the testing of main effects in the analysis of variance. The limited number of power curves restricts the use of the charts to situations in which only a rough estimate of power is required.

Overall and Dalal (7) proposed a method of approximating the power of an F test which is very appealing because of its great simplicity. Their approximation can be denoted as $\hat{F} / F_\alpha$ where $\hat{F}$ is the ratio of the expected mean square between to the expected mean square within and F_α is the critical value of the F ratio with a significance level of α. It should be noted that $\hat{F}$ is not the same as the expected of F since in general the expected value of a ratio is not equal to the ratio of the expected values. Nevertheless, $\hat{F} / F_\alpha$ is very simple to compute and can be readily expressed in terms of the noncentrality parameter λ since

$$\hat{F} = 1 + \lambda . \tag{21}$$

Overall and Dalal have shown that for a particular example the ratio $\hat{F} / F_\alpha$ has a good linear relationship with the true power (correlation equal .988) for a range of true power between .10 and .60. They concluded that $\hat{F} / F_\alpha$ is a good index of power which " . . . should provide an adequate basis for comparing alternative permissible experiments." (7, p. 349). However, for values of the true power less than .10 or greater than .80 the linear fit is not very good. For example, the correlation between true power and $\hat{F} / F_\alpha$ for $\nu_1 = 2$, $\nu_2 = 2, 3, 4, 5, 6, 8, 10, 12, 18, 30$, and 60, and $\lambda = 0, 3, 4, 5, 6, 8, 10, 12$, and 18 is .966 which represents a fit that is considerably less adequate than the one represented by the correlation of .988 reported in the example by Overall and Dalal.

The tabled values of power given by Overall and Dalal (7) and the calculated values of $\hat{F} / F_\alpha$ are presented in columns three and six respectively of Table 3.

Use of the index $\hat{F} / F_\alpha$ in place of the true power can lead to erroneous conclusions about the best allocation of resources. However, the errors will not be serious since an allocation of resources which yields an optimal value of $\hat{F} / F_\alpha$ will yield a true power which will be among the highest possible although it may not be the absolute maximum. In general, $\hat{F} / F_\alpha$ appears to be a useful index: it is easy to calculate and provides a reasonable approximation to power.

The index $\hat{F} / F_\alpha$ does have two minor disadvantages: the obtained values do not have the same scale as power, so the unmodified index does not indicate the actual power level; the index requires only simple hand calculations, but the calculations are based on the tabled values of F_α, so the procedure is not well suited to the computer.

Patnaik (8) has developed an approximation to the noncentral F which fits to the noncentral F, F', a central F distribution with the same first two moments:

$$(22) \quad \int_0^{F'} p_{\nu_1 \nu_2}(F' \mid \lambda)\, dF' \doteq \int_0^{F'/\omega} p_{\nu_1 \nu_2}(F)\, dF$$

where

$$(23) \quad \omega = \frac{\nu_1 + \lambda}{\nu_1}$$

and

$$(24) \quad \nu = \frac{\nu_1 + \lambda)^2}{\nu_1 + 2\lambda} .$$

The accuracy of the approximation appears to be quite good. For those values of power for which Patnaik compares his approximation to the Tang (11) tables, the approximation is generally accurate to two decimal places and the error in the third decimal place appears to be small near the tails. Patnaik's approximation is useful only to the extent that it is possible to evaluate the integral of the appropriate central F distribution. A computer program written by Holloway and Capp provides one method of evaluating the central F integral. This program is presented in Appendix A as Subroutine FDIST.

Table 3

Comparison of Methods of Estimating Power for $\nu_1 = 2$

λ	ν_2	Overall & Dalal Power	Patnaik Approximation	Curve-Fitting Estimate	$\hat{F}/F_\alpha$
0	2	.05	.05	-.04	.05
0	3	.05	.05	-.02	.10
0	4	.05	.05	-.00	.14
0	5	.05	.05	.02	.17
0	6	.05	.05	.04	.20
0	8	.05	.05	.06	.22
0	10	.05	.05	.08	.24
0	12	.05	.05	.10	.26
0	18	.05	.05	.13	.28
0	30	.05	.05	.16	.30
0	60	.05	.05	.15	.32
3	2	.12	.12	.13	.20
3	3	.15	.15	.17	.42
3	4	.18	.18	.19	.58
3	5	.19	.19	.21	.70
3	6	.21	.21	.23	.78
3	8	.23	.23	.25	.90
3	10	.24	.24	.27	.98
3	12	.25	.25	.29	1.03
3	18	.27	.27	.31	1.13
3	30	.28	.28	.34	1.20
3	60	.30	.30	.32	1.27
4	2	.14	.14	.17	.26
4	3	.19	.18	21	.52
4	4	.22	.22	.24	.72
4	5	.24	.24	.26	.86
4	6	.26	.26	.28	.98
4	8	.30	.29	.31	1.12
4	10	.31	.31	.33	1.22
4	12	.33	.33	.34	1.28
4	18	.35	.35	.37	1.41
4	30	.37	.37	.39	1.50
4	60	.39	.39	.38	1.58

Table 3 (Cont'd)

λ	ν_2	Overall & Dalal Power	Patnaik Approximation	Curve-Fitting Estimate	$\hat{F}/F_\alpha$
5	2	.16	.16	.20	.31
5	3	.22	.22	.25	.63
5	4	.26	.26	.28	.86
5	5	.29	.30	.31	1.04
5	6	.32	.32	.33	1.17
5	8	.36	.36	.36	1.34
5	10	.38	.38	.38	1.46
5	12	.40	.40	.40	1.54
5	18	.43	.43	.43	1.69
5	30	.45	.45	.45	1.81
5	60	.47	.47	.43	1.90
6	2	.18	.18	.22	.36
6	3	.25	.25	.28	.74
6	4	.31	.31	.32	1.01
6	5	.35	.35	.35	1.21
6	6	.38	.38	.37	1.36
6	8	.42	.42	.41	1.57
6	10	.45	.45	.43	1.71
6	12	.47	.47	.45	1.80
6	18	.51	.51	.48	1.97
6	30	.54	.54	.51	2.11
6	60	.56	.56	.48	2.22
8	2	.22	.22	.27	.46
8	3	.32	.32	.34	.94
8	4	.39	.39	.39	1.30
8	5	.44	.45	.43	1.56
8	6	.49	.49	.46	1.76
8	8	.54	.54	.50	2.02
8	10	.58	.58	.53	2.20
8	12	.60	.60	.55	2.31
8	18	.65	.64	.59	2.54
8	30	.68	.68	.62	2.71
8	60	.70	.70	.60	2.85

Table 3 (Cont'd)

λ	ν_2	Overall & Dalal Power	Patnaik Approximation	Curve Fitting Estimate	$\hat{F}/F_\alpha$
10	2	.26	.26	.31	.56
10	3	.38	.38	.39	1.16
10	4	.47	.47	.45	1.58
10	5	.53	.54	.50	1.90
10	6	.58	.58	.53	2.14
10	8	.65	.65	.59	2.46
10	10	.68	.68	.62	2.68
10	12	.71	.71	.64	2.83
10	18	.75	.75	.70	3.10
10	30	.79	.79	.73	3.31
10	60	.81	.81	.71	3.49
12	2	.30	.30	.34	.66
12	3	.44	.44	.44	1.36
12	4	.54	.54	.51	1.87
12	5	.61	.62	.57	2.25
12	6	.67	.67	.61	2.54
12	8	.73	.73	.67	2.91
12	10	.77	.77	.71	3.17
12	12	.79	.80	.74	3.34
12	18	.83	.84	.80	3.67
12	30	.86	.86	.84	3.91
12	60	.88	.89	.82	4.12
18	2	.39	.40	.41	.97
18	3	.59	.59	.57	2.00
18	4	.71	.72	.68	2.74
18	5	.79	.79	.76	3.29
18	6	.84	.84	.82	3.70
18	8	.89	.89	.91	4.26
18	10	.92	.92	.97	4.64
18	12	.94	.94	1.02	4.88
18	18	.96	.96	1.09	5.36
18	30	.97	.97	1.15	5.72
18	60	.98	.98	1.15	6.02

The Patnaik approximation and Subroutine FDIST were used to obtain the power estimates reported in column four of Table 3. In only 12 of the 99 power estimates based on the Patnaik approximation in Table 3 is there a difference between these values and the tabled values given by Overall and Dalal (7) as large as .01. Considering that both the tabled values and these estimates have been rounded to the nearest hundredth there is for all practical purposes no difference between the estimates and the tabled values in (7).

It should be noted that the value of ν which was calculated by formula 24 was rounded to the nearest integer before evaluating the integral of the central F distribution. Presumably, the accuracy of the power estimates would be slightly improved by using fractional values of ν, however, in view of the accuracy obtained for the example in Table 3 this may be unnecessary for practical purposes.

In an attempt to determine an easily manipulated function relating power to degrees of freedom and noncentrality parameter, the least-squares method was used to fit power values to functions of the parameters. Primary attention was devoted to the power function for $\nu_1 = 2$. A total of 99 power values were used: the 88 values tabled by Overall and Dalal (7) for $\lambda = 3, 4, 5, 6, 8, 10, 12,$ and 18 and for $\nu_2 = 2, 3, 4, 5, 6, 8, 10, 12, 18, 30,$ and 60; and 11 values of .05 for which $\lambda = 0$ where ν_2 was the same as the tabled values.

For the curve fitting, ϕ^2 and n were substituted for ν_2 and λ:

$$n = \frac{\nu_2}{\nu_1 + 1} + 1 \qquad (25)$$

$$\phi^2 = \lambda/n \; . \qquad (26)$$

Then various functions of n and ϕ^2 were used in the least-squares equations: powers of the parameters ranging from 1/2 to 3, cross-products, and natural logarithms.

The simplest equation with the fewest terms which resulted in the highest correlation with the tabled values was

$$\text{Power} = -10.57 - 1.15n - 8.54\phi^2 + 5.43n\phi^2 + 16.23 \log [n(\phi^2+1)] \; . \qquad (27)$$

The above equation resulted in power estimates that had a correlation of .9812 with the tabled values of Overall and Dalal that are presented in Table 3. These estimates are reported in column five of Table 3. In addition the same equation provided a reasonable fit to the Overall and Dalal power values for $\nu_1 = 1$ ($r = .959$) and $\nu_1 = 1$ and $2 (r = .961)$.

Using this equation for the values for $\nu_1 = 2$, the largest discrepancies between predicted and true values occurred for large values of n and ϕ^2 where the estimated power was greater than one. If a value of 1.0 is substituted for the estimated power values larger than 1.0, the largest discrepancy between predicted and true is .106. This degree of accuracy might be sufficient for some purposes. The accuracy evaluated by the correlation is greater than that of Overall and Dalal's $\hat{F} / F_\alpha$ within the range studied, and the scale is the same as power. However, the computation of the function is far more difficult than $\hat{F} / F_\alpha$, although many values of the function can be quickly computed by a very simple computer program.

Curve fitting as an approach to the power function should not be abandoned. The power functions are not complex curves and there is every reason to believe that a reasonable function can be obtained. Minimizing the squares of the residuals is perhaps not the most appropriate criterion; other criteria should be considered. In addition, future work should use more power values for large n and ϕ^2 so that the asymptote of the power function has better representation.

Cost of an Experiment

It is obvious that an experimenter can always increase power by increasing K and/or n. However, in any practical situation, the experimenter has only limited resources at his command and would like to be able to design the experiment so that the power is maximized within the constraints imposed by the available resources. Generally, the experimenter cannot increase both n and K: if K is increased n must be decreased.

Let C denote the total cost per group of the experiment and assume that this cost is the same for all groups. Following the lead of Cronbach and Gleser (1), it is useful to assume that the cost is the same for all subjects and that this cost per subject consists of a fixed cost, C_o, which is independent of test length and a cost per test unit, C_1. The cost per group, C, is then given by:

$$C = n (C_o + KC_1) \tag{28}$$

where n is the number of people per group and K is the length of the test. Factors which contribute to the fixed cost, C_o, might be length of time required to give instructions and cost of bringing the subject to the testing center. The variable cost, C_1, would be dependent upon factors such as the per-item scoring costs and costs of subject time. There is no real provision in this model for test development costs which would be a function only of K, the test length. This cost model implies that for a constant cost per cell a change in test length, from K to K^* must be accompanied by a change in the number of subjects per cell from n to n^* where

$$n^* = \frac{n\,(C_o + KC_1)}{C_o + K^*C_1} \quad . \tag{29}$$

For any given n, one can solve formula 24 for the maximum allowable K,

$$K = \frac{C - n\,C_o}{C_1} \quad . \tag{30}$$

In the special but rather unrealistic case where C_o is equal to zero, the most efficient allocation of resources will always be achieved by setting K equal to one regardless of the test reliability. This conclusion was drawn by Overall and Dalal (7). This can be seen by noting that for $C_o = 0$, the cost per cell, C, is a constant as long as the product nK is a constant, and for a fixed product, nK, not only is the noncentrality parameter maximized for $K=1$ but so are the degrees of freedom.

Allocation of Resources

To provide the researcher with a method of evaluating the relative effectiveness of increasing sample size and increasing test length, two computer programs were written in FORTRAN IV. Listings of the programs are presented in Appendix A. These programs handle only the limited case of a single factor analysis of variance.

Each program reads six parameters:

1 PHITRU -- the ratio of the variance of the effects to the variance within. This parameter has been denoted above as ϕ_T^2,

2 COST -- the total allowable cost per group, denoted C above,

3 CZERO -- the fixed cost per test, denoted C_o above,

4 CONE -- the variable cost, that is, the cost per test unit, denoted C_1 ,

5 REL -- the reliability of the unit-length test denoted ρ_1 ,

6 VI -- the degrees of freedom for the numerator of the F ratio (number of groups minus one), denoted ν_1 .

Each of the two programs then computes the maximum number of persons per cell permissible within the cost constraints. For each sample size from two to the maximum, the corresponding maximum K is calculated. The λ_X is estimated by using formula 18. Both programs then estimate the power for each of the permissible combinations of n and K . All of the power estimates are printed to permit the identification of the combination of n and K which yields the maximum power.

The first program, "Allocation of resources based on least-squares fit of power function," uses the approximation given by equation 27. This is extremely rapid and can compute power estimates for many combinations in a few seconds. The output of this program consists of the input parameters and K , n , ν_1 ν_2 , ϕ_x^2 , λ_x , and power. Sample computer printouts can be seen in Appendix B.

The second program, "Allocation of resources using the Patnaik approximation," is based upon the noncentral F approximation developed by Patnaik (8) and presented in equations 22, 23, and 24 above. A subroutine FDIST, written by Holloway and Capp and revised by McKelvey (See Appendix A) was used to obtain both the critical F value (F_α) and to evaluate the integral of the central F distribution employed in Patnaik's approximation of the noncentral F distribution. Sample output from this program can also be seen in Appendix B. In addition to the output of the first program, the values of ν in equation 24 and F_α for each permissible combination of n and K are printed. This second program takes significantly more time than the first program: each of the estimated power values requires about four seconds to compute on the IBM 7044.

A comparison of the two methods of estimating power is provided in Table 3. Table 3 also includes the values of power given by Overall and Dalal (7) and the values of $\hat{F} / F_\alpha$, the power approximation suggested by Overall and Dalal (7). As noted before, the scale for $\hat{F} / F_\alpha$ is not the same as scale for power. If a

correlation is used as the measure of the goodness of the approximation, the methods of power can be ordered: $\hat{F} / F_{\alpha}$, $r = .966$; least-squares, $r = .981$; and Patnaik, $r = .9999$. Considering the size of the discrepancies between the estimated and true values, the Patnaik approximation is clearly superior to the least squares.

Table 4 presents the power estimates computed by the two programs under three different cost conditions. The total cost, C, is 3000 $\rho_1 = .10$ and $\nu_1 = 2$ in all cases. Under the first condition $C_o = 0$, and $C_1 = 100$. Under these conditions, power is maximized by increasing sample size to the maximum allowable given the cost constraints, which for all cases represented in Table 4 is 30. The estimates based on the Patnaik approximation accurately reflect this fact. On the other hand, the least-squares estimates erroneously decrease for the largest values of n, but the errors in the estimated power are not large. It is interesting to note that the maximum power in this first, cost condition is much lower than in the other two cases: the large cost per test unit $(C_1 = 100)$ does not permit the use of a very reliable instrument.

In the second and third cost conditions, the maximum power is achieved with a smaller sample size than in the first condition. In these two cases the differences between the allocations based on the two approximations are minimal. However, the differences in the power estimates are not necessarily trivial.

Conclusions

Of the three approximations to the power function that were investigated, the one based on the Patnaik approximation and using the FDIST program to compute integrals of central F distributions was by far the most accurate procedure. The only disadvantage of this method is that it requires considerably more computational time than the other two estimations methods considered.

The least-squares approximation to the power function which was developed has the advantage of great computational speed. However, the method has two major disadvantages in its present state of development; the approximation is limited to the case of two degrees of freedom in the numerator, and the power estimates are not sufficiently accurate for many purposes. In view of the computational ease of this approach, it is considered to be a potentially useful line of future research. If sufficiently accurate estimates could be obtained with a relatively simple function, a major advantage of this approach would be that the function could be dealt with analytically more readily than the integrals of the noncentral F distribution.

Table 4

Estimated Power for C = 3000
$\rho_1 = .10$, $\nu_1 = 2$

	Least-Squares Estimates			Estimates based on Patnaik Approximation		
	C_o	0	80	90	0	80	90
n	C_1	100	20	10	100	20	10
2		.15	.35	.43	.14	.33	.43
3		.22	.52	.69	.19	.55	.74
4		.25	.61	.86	.22	.67	.87
5		.28	.67	.97	.24	.73	.92
6		.30	.70	1.04	.25	.76	.94
7		.31	.72	1.09	.26	.77	.95
8		.32	.73	1.11	.27	.78	.96
9		.33	.73	1.12	.27	.78	.96
10		.33	.72	1.12	.28	.78	.96
11		.33	.71	1.11	.28	.76	.96
12		.34	.70	1.09	.28	.75	.96
13		.34	.69	1.06	.28	.74	.96
14		.34	.67	1.04	.29	.73	.94
15		.34	.65	1.00	.29	.72	.94
16		.33	.63	.97	.29	.71	.93
17		.33	.61	.93	.29	.69	.91
18		.33	.59	.89	.29	.65	.90
19		.33	.56	.84	.29	.63	.89
20		.32	.54	.80	.29	.61	.85
21		.32	.52	.75	.30	.59	.83
22		.31	.49	.70	.30	.57	.78
23		.31	.46	.65	.30	.52	.75
24		.31	.44	.60	.30	.50	.72
25		.30	.41	.54	.30	.47	.65
26		.30	.38	.49	.30	.44	.60
27		.29	.36	.44	.30	.41	.52
28		.29	.33	.38	.30	.36	.46
29		.28	.30	.33	.30	.33	.38
30		.27	.27	.27	.30	.30	.30

[a]The maximum value (based on three decimal places) in each column is underlined.

The Overall and Dalal (7) method of estimating power, $\hat{F} / F_{\alpha}$, is computationally most simple and is the only one of the three methods that is well suited to hand calculations. As previously noted, this approach does not yield the same scale as power, and the estimates are much less accurate than those based on the Patnaik approximation. It would be feasible, of course, to write a computer program which uses a subroutine such as FDIST to compute F_{α}, and then compute and rescale $\hat{F} / F_{\alpha}$ as a means of obtaining power estimates. Such a program presumably would have about twice the speed of the Patnaik approximation program since it would involve only half as many integral evaluations, however, the accuracy of these estimates would not match the accuracy of the Patnaik approximation.

Computer programs were written to determine the most efficient allocation of resources. The two programs are based on the Patnaik and least-squares approximations, and the Patnaik approximation is distinctly superior to the least-squares approximation. It is clear from the sample problems presented that differences in the relative magnitude of fixed and variable cost result in different optimum allocation of resources to test length and sample size. The results are in agreement with the conclusion of Overall and Dalal (7) that the maximum power under conditions of zero fixed cost is always obtained by increasing the sample size to the maximum permissible. Under the more realistic condition of nonzero fixed cost, however, the maximum power is generally obtained with less than maximum permissible sample and corresponding test length which is greater than the minimum unit length test.

Part II: Empirical Demonstration

Purpose

The preceding theoretical development has been based upon the assumptions of classical test theory. Because the assumptions of classical test theory cannot be expected to hold exactly in real data, the effects on the power of statistical tests of changing sample size and test length were demonstrated empirically.

Method

Subjects: The subjects were 4885 eleventh-grade students who had participated in "A Study of Academic Prediction and Growth" a nationwide study sponsored by the Educational Testing Service.

The subjects were divided into groups to permit the study of group comparisons. A two-group division was provided by sex: 2293 males and 2362 females. A three-group division was arbitrarily made by dividing subjects into three groups of approximately equal size on the basis of the mean scores of students in different types of schools. (Schools were divided into nine types for the original study.) The type of school, itself, would have provided a more interesting set of groups for study, but the differences in mean scores were too small. The sizes of the three groups were: low-scoring, 2105; middle-scoring, 1276; high-scoring; 1489. The totals for the two-group and three-group divisions are not the same: subjects with a missing or inappropriate group designation for sex or type of school were eliminated from that analysis.

Measures: In 1961, the subjects responded to 190 verbal type items of the School and College Ability Test (SCAT) and the Sequential Tests of Educational Progress (STEP). These items measure verbal aptitude, reading achievement, and writing achievement. These items were considered to belong to a single item pool.

Procedure: The 190 items were scored to provide each subject with a "true" score. All of the subjects in each of the groups defined above were considered to form a population of interest. The distributions of the true scores were then analyzed for these populations.

To show the effect of the error of measurement on the distribution of the test statistic, items and persons were sampled from the populations according to the scheme presented in Figure 1.

FIGURE 1

Sampling Matrix:

(Number of Samples Drawn for Each Test Length and Sample Size)

Persons Per Group	Items: 10	20	40
	200	200	200
	100	100	100

Tests were created by randomly sampling 10, 20, or 40 items from the total of 190. Samples of persons were created by randomly drawing four or eight persons from each group. For each sample of persons, the items that comprised a single randomly generated test were scored. For each of the designs involving four persons per group, a total of 200 samples were drawn and for each of the designs involving eight persons per group, a total of 100 were drawn. After the randomly generated tests were scored analyses of variance were performed and the distributions of the F statistics were plotted.

Results

Population parameters for each group are presented in Table 5.

TABLE 5

Group True-Score Parameters

		Mean	Standard Deviation	Skewness	Kurtosis
Sex	Male	105.2	33.0	-.335	-.759
	Female	112.6	31.2	-.026	-.933
Scores	Low	99.4	31.0	.125	-.764
	Middle	108.4	32.2	-.100	-.880
	High	122.2	30.5	-.484	-.484

Within each set, the two-group and the three-group, the means are different, the standard deviations comparable, and the measures of

skewness and kurtosis appropriate for the assumption of a normal population distribution. The measures reported for skewness and kurtosis are:

(31) $$\text{Skewness} = \frac{\Sigma (T - \bar{T})^3}{N\sigma_T^2}$$

(32) and $$\text{Kurtosis} = \frac{\Sigma (T - \bar{T})^4}{N\sigma_T^4} - 3$$

where: T is the score on all 190 items for a given subject,
$\bar{T}$ is the group mean, and
σ_T is the group standard deviation.

For a normal population these measure of skewness and kurtosis should be zero.

The results of the analyses of variance are presented in Table 6.

TABLE 6

Population True-Score Analyses of Variance

Source	df	SS	MS	F
Two-Group Analysis				
Between	1	64,025	64,025	62.07
Within	4,653	4,799,809	1,031	
Total	4,654	4,863,834		
Three-Group Analysis				
Between	2	453,338	266,691	232.96
Within	4,867	4,735,485	972	
Total	4,869	5,188,823		

For the two-group analysis (sex) an F ratio of 62.07 with one and 4653 degrees of freedom was obtained. Although in a sample this would obviously be a highly significant F value, the value of ϕ_T^2 is only .0133. Thus the values of λ_T are only .0532 for the designs with four persons per group and .1064 for the designs with eight persons per group.

The F-ratio for the three-group analysis of variance was 232.96 with two and 4867 degrees of freedom. This corresponds to a value of ϕ^2_T equal to .0957. The values of λ_T are .3828 and .7656 for designs with four and eight persons per group respectively.

The distributions of the observed F ratios for the two-group analyses are presented in Figure 2. For two groups and four persons per group there are one and six degrees of freedom and, for $\alpha = .05$ the critical F value is 5.99. With two groups and eight persons per group there are one and 14 degrees of freedom and the critical F value for $\alpha = .05$ is 4.60.

The six distributions shown in Figure 2 do not differ markedly from each other. All six distributions are "J" shaped. For the four-person design there is a steady decrease in the number of F ratios at the low end of the scale as the number of items is increased. In the distributions for eight people the decrease in low values of observed F ratios appears when the number of items increases from 10 to 20 but for 40 items the unusually large number of cases in the lowest interval destroys this trend.

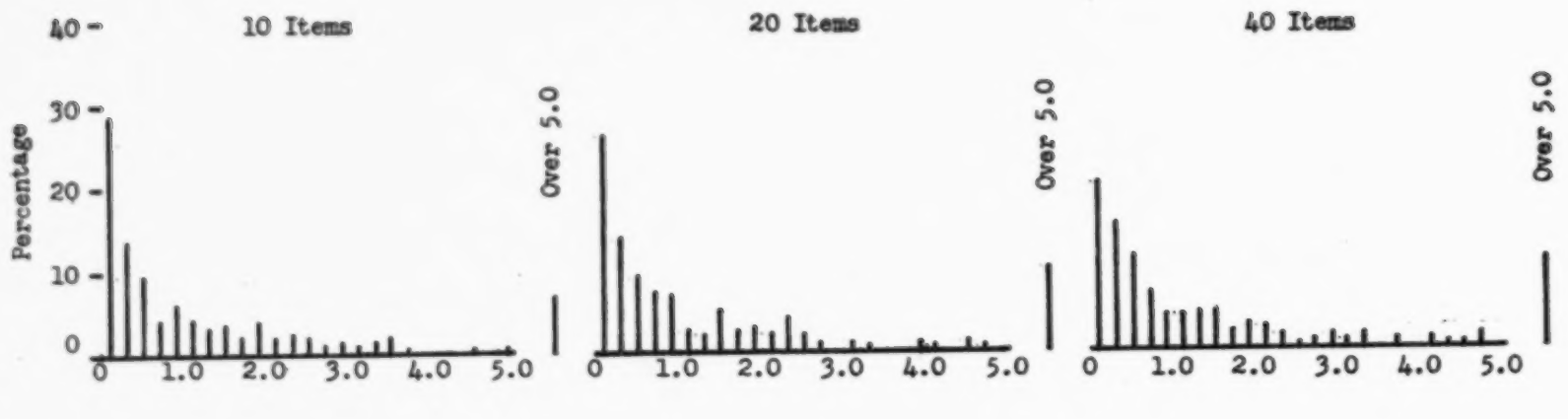
In Figure 3 the comparable distributions of observed F ratios for the three-group analyses are 4.26 and 3.47 for designs with four and eight persons per group respectively. The degrees of freedom for these analyses are two and nine for four persons per group and two and 21 for eight persons per group.

The distributions in Figure 3 are much less "J" shaped, more nearly symmetrical, than their counterparts in Figure 2. The distributions do not change systematically in the four-person designs as the number of items is increased. In the eight-person designs there is some tendency toward larger F ratios as the number of items is increased. The most noticeable difference is between the four- and eight-person designs: larger F ratios are observed in the eight-person designs.

In Table 7, the proportion of observed F ratios that exceed the critical value ($\alpha = .05$) for each experimental design are reported. In all but one of the 12 experimental designs the "observed power" is greater than .05. The observed power is greater in the three-group design than in the two-group design. Within each design there is generally greater observed power for the eight- than for the four-person designs and observed power tends to increase as the number of items is increased.

FIGURE 2

PERCENTAGE DISTRIBUTIONS OF F RATIOS FOR 2 GROUPS



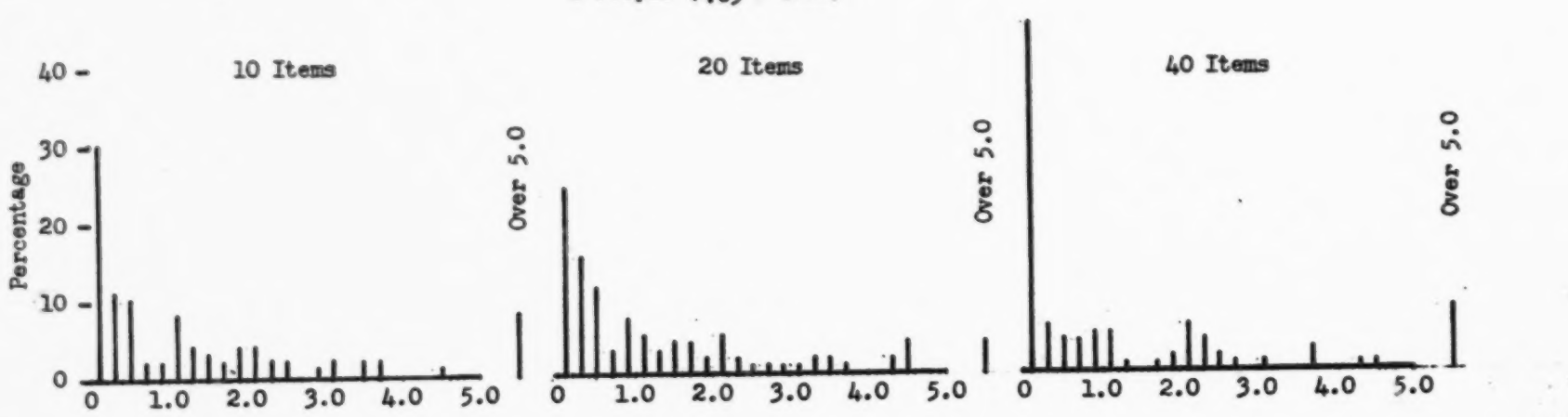
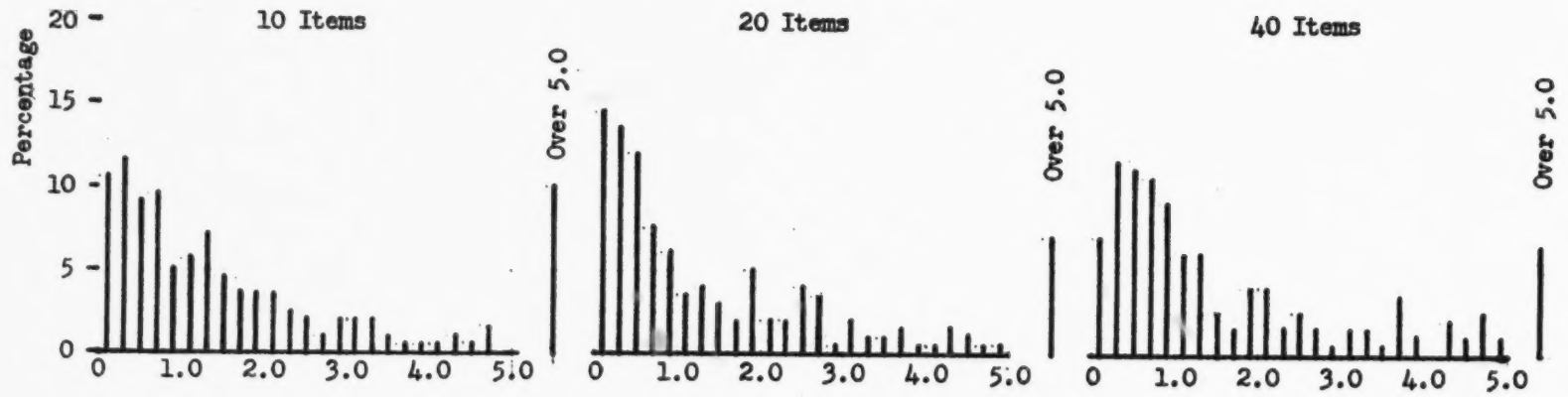


FIGURE 3
PERCENTAGE DISTRIBUTIONS OF F RATIOS FOR 3 GROUPS

4 People ($F_{.05} = 4.26$)



8 People ($F_{.05} = 3.47$)

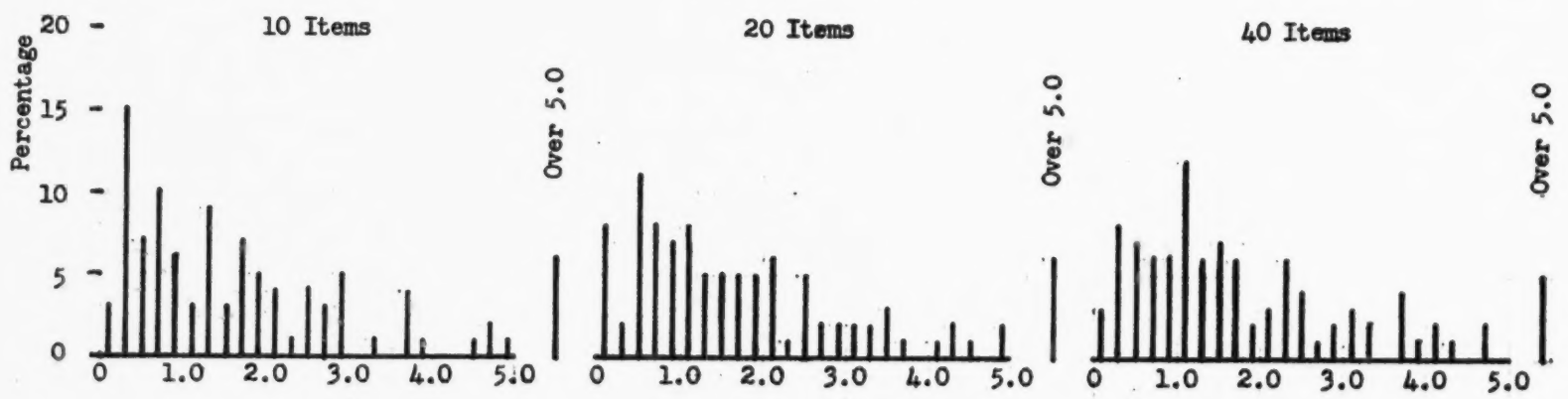


TABLE 7

Observed Power

		Items		
		10	20	40
Two-Group Analyses				
Persons	4	.055	.085	.085
	8	.080	.040	.080
Three-Group Analyses				
Persons	4	.080	.105	.130
	8	.150	.160	.150

Discussion

The empirical distributions of the F ratios presented in Figures 2 and 3 do not contain enough data points to provide very smooth or very stable results. It is clear, however, that the probability of detecting a population true score difference by the methods used is not great.

For the two-group analyses, the population true score non-centrality parameters, λ_T, are only .0532 and .1064 for the four- and eight-person analyses respectively. The values of λ_X are even smaller. The variance of the group effects, σ^2_A, is in each case 13.75. Relative to the within group variance of 1031, this variance is very small and in view of the λ_T values one would not expect the power to be much greater than .05 and it was not.

For the three-group analyses, the population true score non-centrality parameters, λ_T, are .3828 and .7656 for the four- and eight-person analyses respectively. While not large, these values are on the order of eight times as large as the corresponding λ_T values in the two-group analyses and the degrees of freedom for both numerator and denominator are larger for the three-group analyses

than they are for the two-group analyses. The variance of the group effects, σ_A^2, is in each case, 93.01. The theoretical power for the three-group analyses would be greater than for the two-group analyses, but it would still be less than .10. The observed power which is reported in Table 7 was greater for the three-group analyses than for the two-group analyses.

Although the true score population differences were small in both examples, the differences could be of psychological or educational importance. But it is clear that differences of this magnitude are not likely to be detected with samples of the size used in these examples.

Conclusions

The empirical demonstration of the effect of error of measurement of the power of statistical tests was limited by the small population true score differences among the groups. The expected trends were not clearly demonstrated but some indication of increasing power with increasing reliability of the instrument and with increasing sample size were observed. The effect of increasing n appeared to be relatively greater than the effect of increasing K which is in agreement with theoretical expectations.

The most striking feature of the demonstration, however, was that the population mean differences which are reported in Table 4 and which appear to reflect the magnitude of differences in which the educational researcher is often interested have little chance of being detected with the four- or eight-person designs studied here.

Summary

The purpose of this research was to study the effect of error of measurement upon the power of statistical tests. Attention was focused on the F test of the single factor analysis of variance. Formulas were derived to show the relationship between the noncentrality parameters for analyses using true scores and those using observed scores. The effect of the reliability of the measurement and the sample size were thus demonstrated. The assumptions of classical test theory were used to develop formulas relating test length to the noncentrality parameters.

Three methods of estimating power for different conditions of sample size and test length were studied. The three methods were: $\hat{F} / F_{\alpha}$ suggested by Overall and Dalal (7), a least-squares approximation, and an approximation based on the work of Patnaik (8). The approximation based on Patnaik's work was significantly more accurate than the other two methods but required more computational time.

The cost of an experiment was analyzed in terms of a fixed cost per subject and a variable cost dependent upon test length. Computer programs were written to use the least-squares approximation and the approximation based on Patnaik to estimate the power under all permissible allocations of resources to sample size and test length. The program results indicate which of the permissible allocations will result in maximum power.

To demonstrate empirically the effect of error of measurement on the power of statistical tests, samples of persons and items were randomly drawn from a large pool of data. Tests of 10, 20, and 40 randomly drawn items were scored for samples with four- and eight-persons per group. The expected trends were present but not definitive.

References

1. Cronbach, L. J.; and Gleser, Goldine C. Psychological Tests and Personnel Decisions. (2nd ed.) Urbana, Illinois: University of Illinois Press. 1965.

2. Feldt, L. S., and Mahmoud, M. W. "Power Function Charts for Specification of Sample Size in Analysis of Variance," Psychometrika. XXIII, September 1958. p. 201-210.

3. Fox, M. "Charts for the Power of the F-Test." Annals of Mathematical Statistics. XXVII, June 1956. p. 484-497.

4. Gulliksen, H. Theory of Mental Tests. New York: John Wiley & Sons, Inc. 1950.

5. Lehmer, Emma. "Inverse Tables of Probabilities of Errors of the Second Kind." Annals of Mathematical Statistics. XV, December 1944. p. 388-398.

6. Lord, F. M. Item Sampling in Test Theory and in Research Design. Research Bulletin 65-22, Princeton, N. J., Educational Testing Service, 1965.

7. Overall, J. E., and Dalal, S. N. "Design of Experiments to Maximize Power Relative to Cost," Psychological Bulletin. LXIV, November 1965. p. 339-350.

8. Patnaik, P. B. "The Non-Central χ^2- and F-Distributions and Their Applications," Biometrika. XXXVI, June 1949. p. 202-232.

9. Pearson, E. S., and Hartley, H. O. "Charts of the Power Function for Analysis of Variance Tests, Derived from the Non-Central F-Distribution," Biometrika. XXXVIII, June 1951. p. 112-130.

10. Sutcliffe, J. P. "Error of Measurement and the Sensitivity of a Test of Significance," Psychometrika. XXIII, March 1958. p. 9-17.

11. Tang, P. C. "The Power Function of the Analysis of Variance Test," Statistical Research Memoirs. II, 1938. p. 126-149.

12. Wishart, J. "A Note on the Distribution of the Correlation Ratio," Biometrika. XXIV, November 1932. p. 441-456.

APPENDIX A

FORTRAN Program Lists

Program	page
Program using least-squares approach	A-1
Program based on Patnaik approximation	A-2
Subroutine FDIST	A-3

```
C          ALLOCATION OF RESOURCES BASED ON LEAST SQUARES FIT OF POWER
C          FUNCTION

    1 CONTINUE
      READ 5, PHITRU, COST, CZERO, CONE,REL                    ,V1
    5 FORMAT(   F5.1, 3F5.0, F5.2, F5.0)
      IF(PHITRU) 99,99,88
   88 CONTINUE
      PRINT 6, PHITRU, COST, CZERO, CONE ,REL     ,V1
    6 FORMAT(1H1, F5.1, 3F7.0, F5.2, F5.0)
      PRINT 7,
    7 FORMAT(99H0          K          N        NU1        NU2 OB PHI**2 OBS
     1LAMDA     POWER                                        )
      NMAX = COST/( CZERO + CONE)
      DO 20 N= 2,NMAX
      PEOP = N
      XK = (COST - PEOP*CZERO)/(PEOP*CONE)
      PHIOB = (XK * REL*PHITRU)/ (REL*XK + (1.0-REL)*(PHITRU+1.0))
      POWER = -10.57 -1.15*PEOP -8.54*PHIOB  +5.43*PHIOB*PEOP +
     1 16.23*ALOG(PEOP*(PHIOB+1.0))
      POWER = POWER/100.
      XLAM = PEOP*PHIOB
      V2 = 3*( N-1)
      IV1 = V1
      IV2 = V2
      PRINT 10, XK,N, IV1, IV2, PHIOB, XLAM    ,POWER
   10 FORMAT (F11.3, 3I10, 4F10.3)
      PUNCH 11 ,XK,N,IV1,IV2,PHIOB,XLAM,POWER
   11 FORMAT (F8.3,3I6,4F8.3)
   20 CONTINUE
      GO TO 1
   99 CONTINUE
      END
```

```
C      ALLOCATION OF RESOURCES USING THE PATNAIK APPROXIMATION

    1 CONTINUE
      READ 5, PHITRU, COST, CZERO, CONE,REL ,V1
    5 FORMAT(   F5.1, 3F5.0, F5.2, F5.0)
      IF(PHITRU) 99,99,88
   88 CONTINUE
      PRINT 6, PHITRU, COST, CZERO, CONE ,REL
    6 FORMAT(1H1, F5.1, 3F7.0, F5.2)
      PRINT 7,
    7 FORMAT (90H0        K           N        NU1        NU2          NU  OBS
     1PHI**2 OBS LAMDA  FALPHA       POWER       )
      NMAX = COST/( CZERO + CONE)
      DO 20 N= 2,NMAX
      PEOP = N
      XK = (COST - PEOP*CZERO)/(PEOP*CONE)
      PHIOB = (XK * REL*PHITRU)/ (REL*XK + (1.0-REL)*(PHITRU+1.0))
      XLAM = PEOP*PHIOB
      V2 = 3*( N-1)
      PHI = SQRT (PHIOB)
      IV2 = V2
      FALPHA = 0.
      CALL FDIST(2,IV2,FALPHA,.95)
      GALPHA = FALPHA
      PHIOB = PEOP*PHIOB
      SCALE = (2.+PHIOB)/ 2.
      FALPHA = FALPHA/ SCALE
      V= (( 2. + PHIOB)**2)/ (2.+2. *PHIOB)
      V = V+ .5
      IV = V
      PROB = 0.
      CALL FDIST (IV,IV2, FALPHA, PROB)
      PROB = 1.0 - PROB
      IV1 = V1
      IV2 = V2
      XYLAM = PHIOB
      PHIOB = PHIOB/PEOP
      PRINT 70, XK,N, IV1, IV2, V, PHIOB, XYLAM, GALPHA, PROB
   70 FORMAT (F10.3,3I10,5F10.3)
      PUNCH 71,XK,N,IV1,IV2,V,PHIOB,XYLAM,GALPHA,PROB
   71 FORMAT (F8.3,3I6,5F8.3)
   20 CONTINUE
      GO TO 1
   99 CONTINUE
      END
```

```
      SUBROUTINE FDIST  (MM,NN,FX,PROBX)

C     CLARK HOLLOWAY AND W.B.CAPP, AUGUST 31,1959
C     REVISED APRIL 1,1961   R.J.MCKELVEY
      DIMENSION B(2)
      NOUT=6
      SF    = 0.0
      SPROB =0.0
      F=FX
      PROB=PROBX
      M=MM
      N=NN
      IF(F) 76,106,100
  100 SF= F
      IF (F-1.0) 101, 101,105
  101 XM=M
      XN= N
      LOW = 1
      FLO = F
  102 PLO = 0.0
      DELTA=FLO/500.0
      GO TO 21
  105 FLO=1.0/F
      XM= N
      XN= M
      LOW = 0
      GO TO 102
  106 SPROB = PROB
      IF(PROB)76,76,107
  107 IF (PROB- 0.5) 108,108,110
  108 XM=M
      XN=N
      LOW= 1
      PLO = PROB
  109 FLO = 0.0
      DELTA=PLO/200.0
      GO TO 21
  110 IF(PROB-1.)111,76,76
  111 XM=N
      XN=M
      LOW = 0
      PLO = 1.0 - PROB
      GO TO 109
   21 FACTL=0.0
  215 FACT=1.0
      B(1)=(XM-2.0)/2.0
      B(2)=(XN-2.0)/2.0
   24 A=(XM+XN-2.0)/2.0
```

```
241   IF(A-0.2)400,76,242
400   FACT = 0.31830989
      GO TO 283
242   IF(A-0.7)410,76,243
410   FACT=0.5
      GO TO 283
243   DO 245 I=1,2
      IF(B(I)-0.7)261,76,245
261   IF(B(I)-0.2)264,76,262
262   FACT=FACT/0.886226925
263   B(I)=1.0
      GO TO 245
264   IF(B(I)+0.2)265,76,263
265   FACT=FACT/1.772453850
      GO TO 263
245   CONTINUE
244   IF(A-0.7)281,76,251
251   FACT=FACT*A/(B(1)*B(2))
      IF(FACT-99999999.)830,283,283
 830  IF(FACT-1.0E-8)283,283,26
26    A=A-1.0
      B(1)=B(1)-1.0
      B(2)=B(2)-1.0
      GO TO 243
281   IF(A-0.2)283,76,282
282   FACT=FACT*0.886226925
283   FACTL=FACTL+ALOG(FACT)
      FACT=1.0
      IF(A-0.7)284,76,26
284   Y1=FACTL+(XM/2.0)*ALOG(XM/XN)
      Y2=(XM-2.0)/2.0
      Y3=(XM+XN)/2.0
36    F=DELTA/2.0
      CUM=0.0
C
37    HFDL=Y1+Y2*ALOG(F)-Y3*ALOG(1.0+XM*F/XN)+ALOG(DELTA)
      IF(HFDL+20.)50,51,51
50    HFD=0.0
      GO TO 52
51    HFD=EXP (HFDL)
52    CUM=CUM+HFD
      F=F+DELTA
375   IF(F-FLO)37,37,38
38    IF(PLO)76,39,381
381   IF(HFD)76,384,382
382   IF(ALOG(PLO)-HFDL-4.604)383,384,384
383   DELTA=DELTA/2.0
      GO TO 36
```

```
384   IF(CUM-PLO)37,39,39
39    FLO=F-DELTA
      IF(SF)    76,43,40
40    F = SF
      IF(LOW) 76,42,41
41    PROB = CUM
      GO TO 49
42    PROB = 1.0- CUM
      GO TO 49
43    PROB = SPROB
      IF(LOW) 76,45,44
44    F = FLO
      GO TO 49
45    F = 1.0/FLO
49    PROBX=PROB
      FX=F
1000  RETURN
  76  WRITE (6,176)              MM,MN,FX,PROBX
 176  FORMAT (10X,36HCOULD NOT WORK F DISTRIBUTION WITH
     1(I6,1H,I6,1H,E131.6,1H,E13.6,1H))
      GO TO 1000
      END
```

APPENDIX B

Sample Program Output

Source	page
Program using least-squares approach	B-1
Program based on Patnaik approximation	B-4

100 3000 0 100 10 2

K	N	NU1	NU2	OBPHI**2	OB LAMDA	POWER
15.000	2	2	3	1.316	2.632	0.151
10.000	3	2	6	0.917	2.752	0.215
7.500	4	2	9	0.704	2.817	0.253
6.000	5	2	12	0.571	2.857	0.278
5.000	6	2	15	0.481	2.885	0.295
4.286	7	2	18	0.415	2.905	0.308
3.750	8	2	21	0.365	2.920	0.318
3.333	9	2	24	0.326	2.932	0.325
3.000	10	2	27	0.294	2.941	0.329
2.727	11	2	30	0.268	2.949	0.333
2.500	12	2	33	0.246	2.956	0.335
2.308	13	2	36	0.228	2.961	0.336
2.143	14	2	39	0.212	2.966	0.336
2.000	15	2	42	0.198	2.970	0.335
1.875	16	2	45	0.186	2.974	0.334
1.765	17	2	48	0.175	2.977	0.332
1.667	18	2	51	0.166	2.980	0.329
1.579	19	2	54	0.157	2.983	0.326
1.500	20	2	57	0.149	2.985	0.322
1.429	21	2	60	0.142	2.987	0.319
1.364	22	2	63	0.136	2.989	0.314
1.304	23	2	66	0.130	2.991	0.310
1.250	24	2	69	0.125	2.993	0.305
1.200	25	2	72	0.120	2.994	0.300
1.154	26	2	75	0.115	2.995	0.295
1.111	27	2	78	0.111	2.997	0.289
1.071	28	2	81	0.107	2.998	0.283
1.034	29	2	84	0.103	2.999	0.277
1.000	30	2	87	0.100	3.000	0.271

100 3000 80 20 10 2

K	N	NU1	NU2	OBPHI**2	OB LAMDA	POWER
71.000	2	2	3	4.176	8.353	0.348
46.000	3	2	6	3.172	9.517	0.516
33.500	4	2	9	2.528	10.113	0.611
26.000	5	2	12	2.080	10.400	0.668
21.000	6	2	15	1.750	10.500	0.701
17.429	7	2	18	1.497	10.479	0.719
14.750	8	2	21	1.297	10.374	0.727
12.667	9	2	24	1.134	10.209	0.728
11.000	10	2	27	1.000	10.000	0.723
9.636	11	2	30	0.887	9.757	0.714
8.500	12	2	33	0.791	9.488	0.702
7.538	13	2	36	0.708	9.199	0.687
6.714	14	2	39	0.635	8.892	0.670
6.000	15	2	42	0.571	8.571	0.651
5.375	16	2	45	0.515	8.240	0.631
4.824	17	2	48	0.465	7.898	0.610
4.333	18	2	51	0.419	7.548	0.587
3.895	19	2	54	0.379	7.192	0.564
3.500	20	2	57	0.341	6.829	0.540
3.143	21	2	60	0.308	6.462	0.515
2.818	22	2	63	0.277	6.089	0.490
2.522	23	2	66	0.248	5.713	0.464
2.250	24	2	69	0.222	5.333	0.437
2.000	25	2	72	0.198	4.950	0.410
1.769	26	2	75	0.176	4.565	0.383
1.556	27	2	78	0.155	4.177	0.356
1.357	28	2	81	0.135	3.786	0.328
1.172	29	2	84	0.117	3.394	0.300
1.000	30	2	87	0.100	3.000	0.271

100 3000 90 10 10 2

K	N	NU1	NU2	OBPHI**2	OB LAMDA	POWER
141.000	2	2	3	5.875	11.750	0.433
91.000	3	2	6	4.789	14.368	0.694
66.000	4	2	9	4.000	16.000	0.862
51.000	5	2	12	3.400	17.000	0.971
41.000	6	2	15	2.929	17.571	1.042
33.857	7	2	18	2.548	17.839	1.086
28.500	8	2	21	2.235	17.882	1.110
24.333	9	2	24	1.973	17.757	1.120
21.000	10	2	27	1.750	17.500	1.118
18.273	11	2	30	1.558	17.140	1.107
16.000	12	2	33	1.391	16.696	1.089
14.077	13	2	36	1.245	16.184	1.065
12.429	14	2	39	1.115	15.615	1.036
11.000	15	2	42	1.000	15.000	1.003
9.750	16	2	45	0.897	14.345	0.967
8.647	17	2	48	0.803	13.656	0.927
7.667	18	2	51	0.719	12.937	0.885
6.789	19	2	54	0.642	12.194	0.841
6.000	20	2	57	0.571	11.429	0.796
5.286	21	2	60	0.507	10.644	0.748
4.636	22	2	63	0.447	9.842	0.699
4.043	23	2	66	0.392	9.025	0.649
3.500	24	2	69	0.341	8.195	0.598
3.000	25	2	72	0.294	7.353	0.545
2.538	26	2	75	0.250	6.500	0.492
2.111	27	2	78	0.209	5.637	0.438
1.714	28	2	81	0.170	4.766	0.383
1.345	29	2	84	0.134	3.887	0.327
1.000	30	2	87	0.100	3.000	0.271

100 3000 0 100 10 2

K	N	NU1	NU2	NU	OBPHI**2	OB LAMDA	FALPHA	POWER
15.000	2	2	3	3.453	1.316	2.632	9.558	0.137
10.000	3	2	6	3.509	0.917	2.752	5.145	0.193
7.500	4	2	9	3.539	0.704	2.817	4.258	0.223
6.000	5	2	12	3.558	0.571	2.857	3.885	0.241
5.000	6	2	15	3.571	0.481	2.885	3.682	0.253
4.286	7	2	18	3.580	0.415	2.905	3.554	0.262
3.750	8	2	21	3.587	0.365	2.920	3.468	0.268
3.333	9	2	24	3.593	0.326	2.932	3.403	0.273
3.000	10	2	27	3.597	0.294	2.941	3.354	0.277
2.727	11	2	30	3.601	0.268	2.949	3.315	0.280
2.500	12	2	33	3.604	0.246	2.956	3.285	0.283
2.308	13	2	36	3.607	0.228	2.961	3.259	0.285
2.143	14	2	39	3.609	0.212	2.966	3.238	0.287
2.000	15	2	42	3.611	0.198	2.970	3.219	0.289
1.875	16	2	45	3.613	0.186	2.974	3.204	0.290
1.765	17	2	48	3.614	0.175	2.977	3.191	0.291
1.667	18	2	51	3.616	0.166	2.980	3.178	0.292
1.579	19	2	54	3.617	0.157	2.983	3.168	0.293
1.500	20	2	57	3.618	0.149	2.985	3.158	0.294
1.429	21	2	60	3.619	0.142	2.987	3.151	0.295
1.364	22	2	63	3.620	0.136	2.989	3.143	0.296
1.304	23	2	66	3.621	0.130	2.991	3.136	0.296
1.250	24	2	69	3.621	0.125	2.993	3.129	0.297
1.200	25	2	72	3.622	0.120	2.994	3.124	0.298
1.154	26	2	75	3.623	0.115	2.995	3.119	0.298
1.111	27	2	78	3.623	0.111	2.997	3.114	0.299
1.071	28	2	81	3.624	0.107	2.998	3.109	0.299
1.034	29	2	84	3.625	0.103	2.999	3.104	0.300
1.000	30	2	87	3.625	0.100	3.000	3.102	0.300

100 3000 80 20 10 2

K	N	NU1	NU2	NU	OBPHI**2	OB LAMDA	FALPHA	POWER
71.000	2	2	3	6.230	4.176	8.353	9.558	0.329
46.000	3	2	6	6.806	3.172	9.517	5.145	0.553
33.500	4	2	9	7.102	2.528	10.113	4.258	0.672
26.000	5	2	12	7.244	2.080	10.400	3.885	0.726
21.000	6	2	15	7.293	1.750	10.500	3.682	0.755
17.429	7	2	18	7.283	1.497	10.479	3.554	0.771
14.750	8	2	21	7.231	1.297	10.374	3.468	0.779
12.667	9	2	24	7.149	1.134	10.209	3.403	0.782
11.000	10	2	27	7.045	1.000	10.000	3.354	0.782
9.636	11	2	30	6.925	0.887	9.757	3.315	0.755
8.500	12	2	33	6.792	0.791	9.488	3.285	0.750
7.538	13	2	36	6.648	0.708	9.199	3.259	0.742
6.714	14	2	39	6.496	0.635	8.892	3.238	0.733
6.000	15	2	42	6.338	0.571	8.571	3.219	0.722
5.375	16	2	45	6.174	0.515	8.240	3.204	0.709
4.824	17	2	48	6.005	0.465	7.898	3.191	0.694
4.333	18	2	51	5.833	0.419	7.548	3.178	0.651
3.895	19	2	54	5.657	0.379	7.192	3.168	0.634
3.500	20	2	57	5.478	0.341	6.829	3.158	0.614
3.143	21	2	60	5.298	0.308	6.462	3.151	0.593
2.818	22	2	63	5.115	0.277	6.089	3.143	0.570
2.522	23	2	66	4.931	0.248	5.713	3.136	0.521
2.250	24	2	69	4.746	0.222	5.333	3.129	0.496
2.000	25	2	72	4.559	0.198	4.950	3.124	0.469
1.769	26	2	75	4.372	0.176	4.565	3.119	0.440
1.556	27	2	78	4.185	0.155	4.177	3.114	0.408
1.357	28	2	81	3.998	0.135	3.786	3.109	0.365
1.172	29	2	84	3.811	0.117	3.394	3.104	0.333
1.000	30	2	87	3.625	0.100	3.000	3.102	0.300

100 3000 90 10 10 2

K	N	NU1	NU2	NU	OBPHI**2	OB LAMDA	FALPHA	POWER
141.000	2	2	3	7.914	5.875	11.750	9.558	0.429
91.000	3	2	6	9.217	4.789	14.368	5.145	0.745
66.000	4	2	9	10.029	4.000	16.000	4.258	0.870
51.000	5	2	12	10.528	3.400	17.000	3.885	0.917
41.000	6	2	15	10.813	2.929	17.571	3.682	0.938
33.857	7	2	18	10.946	2.548	17.839	3.554	0.950
28.500	8	2	21	10.968	2.235	17.882	3.468	0.956
24.333	9	2	24	10.905	1.973	17.757	3.403	0.959
21.000	10	2	27	10.777	1.750	17.500	3.354	0.960
18.273	11	2	30	10.597	1.558	17.140	3.315	0.960
16.000	12	2	33	10.376	1.391	16.696	3.285	0.959
14.077	13	2	36	10.121	1.245	16.184	3.259	0.957
12.429	14	2	39	9.838	1.115	15.615	3.238	0.944
11.000	15	2	42	9.531	1.000	15.000	3.219	0.939
9.750	16	2	45	9.205	0.897	14.345	3.204	0.933
8.647	17	2	48	8.862	0.803	13.656	3.191	0.911
7.667	18	2	51	8.505	0.719	12.937	3.178	0.900
6.789	19	2	54	8.135	0.642	12.194	3.168	0.888
6.000	20	2	57	7.755	0.571	11.429	3.158	0.852
5.286	21	2	60	7.365	0.507	10.644	3.151	0.832
4.636	22	2	63	6.967	0.447	9.842	3.143	0.783
4.043	23	2	66	6.563	0.392	9.025	3.136	0.754
3.500	24	2	69	6.152	0.341	8.195	3.129	0.719
3.000	25	2	72	5.736	0.294	7.353	3.124	0.649
2.538	26	2	75	5.317	0.250	6.500	3.119	0.600
2.111	27	2	78	4.894	0.209	5.637	3.114	0.519
1.714	28	2	81	4.470	0.170	4.766	3.109	0.457
1.345	29	2	84	4.046	0.134	3.887	3.104	0.384
1.000	30	2	87	3.625	0.100	3.000	3.102	0.300

OE 6000 (Rev. 9-66)

INSTRUCTIONS FOR COMPLETING ERIC REPORT RESUME

The resume is used to identify summary data and information about each document acquired, processed, and stored within the ERIC system. In addition to serving as a permanent record of the document in the collection, the resume is also a means of dissemination. All fields of the form must be completed in the allotted spaces, but inapplicable fields should be left blank. The following instructions are keyed to the line numbers appearing in the left margin of the form:

TOP LINE. ERIC Accession No. Leave blank. A permanent ED number will be assigned to each resume and its corresponding document as they are processed into the ERIC system.

LINE 001. Clearinghouse Accession No. For use only by ERIC Clearinghouses. Enter the alpha code and 6-digit document number.

Resume Date. In numeric form, enter month, day, and year that resume is completed. *(Example: 07 14 68.)*

F.A. Leave blank.

T.A. Leave blank.

Copyright. Check appropriate block to denote presence of copyrighted material within the document.

ERIC Reproduction Release. Check appropriate block to indicate that ERIC has permission to reproduce the document and its resume form.

LINES 100–103. Title. Enter the complete document title, including subtitles if they add significant information. Where applicable, also enter volume number or part number, and the type of document *(Final Report, Interim Report, Thesis, etc.).*

LINE 200. Personal Author(s). Enter personal author(s), last name first. *(Example: Doe, John J.)* If two authors are given, enter both. *(Example: Doe, John J.; Smith, Ted).* If there are three or more authors, list only one followed by "and others."

LINE 300. Institution *(Source).* Enter the name of the organization which originated the report. Include the address *(city and State)* and the subordinate unit of the organization. *(Example: Harvard Univ., Cambridge, Mass., School of Education.)*

Source Code. Leave blank.

LINE 310. Report/Series No. Enter any unique number assigned to the document by the institutional source. *(Example: EC-1234)*

LINE 320. Other Source. Use only when a second source is associated with the document. Follow instructions for Line 300 above.

Source Code. Leave blank.

LINE 330. Other Report No. Enter document number assigned by the second source.

LINE 340. Other Source. Use only when a third source is associated with the document. Follow instructions for Line 300 above.

Source Code. Leave blank.

LINE 350. Other Report No. Enter document number assigned by the third source.

LINE 400. Publication Date. Enter the day, month, and year of the document. *(Example: 12 Jun 68.)*

Contract/Grant Number. Applicable only for documents generated from research sponsored by the U.S. Office of Education. Enter appropriate contract or grant number and its prefix. *(Example: OEC-1-6-061234-0000)*

LINES 500–501. Pagination, etc. Enter the total number of pages of the document, including illustrations and appendixes. *(Example: 115p.)* USE THIS SPACE FOR ADDITIONAL INFORMATION PERTINENT TO THE DOCUMENT, such as publisher, journal citation, and other contract numbers.

LINES 600–606. Retrieval Terms. Enter the important subject terms *(descriptors)* which, taken as a group, adequately describe the contents of the document.

LINE 607. Identifiers. Enter any additional important terms, more specific than descriptors, such as trade names, equipment model names and numbers, organization and project names, discovered in the document.

LINES 800–822. Abstract. Enter an informative abstract of the document. Its style and content must be suitable for public announcement and dissemination.

U.S. GOVERNMENT PRINTING OFFICE: 1968 O—331-830

OE 6000 (rev. 9-66)

DEPARTMENT OF HEALTH, EDUCATION AND WELFARE
OFFICE OF EDUCATION

(TOP)

ERIC REPORT RESUME

ERIC ACCESSION NO.

REPORT/ACCESSION NUMBER

PUBLICATION DATE 09-27-67

P.E.

P.E.

REPRODUCTION RELEASE? YES ☒ NO ☐

ERIC REPRODUCTION RELEASE? YES ☒ NO ☐

100 TITLE

EFFECT OF ERROR OF MEASUREMENT ON THE POWER OF STATISTICAL TESTS.

Final Report

200 PERSONAL AUTHOR(S)

Cleary, T. Anne and Linn, Robert L.

300 INSTITUTION (SOURCE)

Educational Testing Service, Princeton, New Jersey

SOURCE CODE

310 INSTITUTION (SOURCE)

320 SPONSORING AGENCY

University of Wisconsin, Madison, Wisconsin

SOURCE CODE

330 SPONSORING AGENCY

340 REPORT/SERIES NO.

SOURCE CODE

350 REPORT/SERIES NO.

400 PUB. DATE 09-27-67 CONTRACT/GRANT NUMBER

500 PAGINATION, ETC.

i + 43 pp.

501

600 KEY WORDS

601 Statistics

602 Mental Test Theory

603 Power

604 Reliability

605 Error of Measurement

606

607 IDENTIFIERS

608 ABSTRACT

609 The purpose of this research was to study the effect of error of measure-
610 ment upon the power of statistical tests. Attention was focused on the F test
611 of the single factor analysis of variance. Formulas were derived to show the
612 relationship between the noncentrality parameters for analyses using true scores
613 and those using observed scores. The effect of the reliability of the measure-
614 ment and the sample size were thus demonstrated. The assumptions of classical
615 test theory were used to develop formulas relating test length to the non-
616 centrality parameters.

617 Three methods of estimating power for different conditions of sample size
618 and test length were studied. The cost of an experiment was analyzed in terms
619 of a fixed cost per subject and a variable cost dependent upon test length.
620 Computer programs were written to use the least squares approximation and the
621 approximation based on Fatnaik to estimate the power under all permissible
622 allocations of resources to sample size and test length. The program results
indicate which of the permissible allocations will result in maximum power.

To demonstrate empirically the effect of error of measurement on the power of statistical tests, samples of persons and items were randomly drawn from a large pool of data. Tests of 10, 20, and 40 randomly drawn items were scored for samples with four and eight persons per group. The expected trends were present but not definitive.